2010

Testing Models or Fitting Models? Identifying Model Misspecification in PLS

Joerg Evermann
Memorial University of Newfoundland St. John's, jevermann@mun.ca

Mary Tate
Victoria University of Wellington, Mary.tate@vuw.ac.nz

Testing Models or Fitting Models? Identifying Model Misspecification in PLS

Completed Research Paper

Joerg Evermann
Memorial University of Newfoundland
St. John's, Canada
jevermann@mun.ca

Mary Tate
Victoria University of Wellington
Wellington, New Zealand
mary.tate@vuw.ac.nz

Abstract

Partial Least Squares (PLS) is a statistical technique that is widely used in the Information Systems discipline to estimate statistical models with structural equations and latent variables. While PLS does not provide a statistical test of model fit to data, its proponents have suggested a set of criteria that good PLS models should fulfill. Conversely, when a model does not satisfy these criteria, it would be judged a bad model. In this paper, we report on the results of a simulation study to examine to what extent the proposed model quality criteria are able to identify misspecified models.

Keywords: Partial Least Squares, Misspecification, Simulation Study, Structural Equation Modeling, Data analysis, Research methods

Introduction

In Information Systems (IS) research, structural equation models are analyzed using either covariance-structure analysis (CB-SEM) or partial least squares analysis (PLS). A recent study identifies IS as the primary user of PLS analysis (Rouse and Corbitt, 2008) and a quick survey of four IS journals (MISQ, ISR, JMIS, and JAIS) from 2004 through 2008 identified 76 studies using PLS, 54 studies using CB-SEM, and 6 studies using both PLS and CB-SEM.

While PLS and CB-SEM are based on fundamentally different principles, both allow the researcher to fit a parameterized model to data and obtain best estimates for the model parameters. For CB-SEM, the notion of "best" is based on explaining observed covariances in the data, while for PLS, the notion of "best" is based on explaining observed variance in the endogenous latent variables (Gefen *et al.,* 2000). The parameter estimates of the two methods converge with increasing number of indicator variables and sample size (Lohmöller, 1989). Both techniques allow the estimated parameters to be tested for statistical significance. PLS uses bootstrap re-sampling, while CB-SEM uses a test statistic based on standard errors.

On closer examination, the logic of statistical testing and fitting differs between CB-SEM and PLS. In CB-SEM, model-implied covariances are statistically tested for differences from the observed covariances, and model parameters are fitted to the data to minimize those differences. The subsequent parameter significance test therefore examines $p(\theta \mid M)$ where θ is the set of fitted parameters and M is the *accepted* model. In PLS, there is no such overall model test. The parameter significance test also examines $p(\theta \mid M)$ but M is the *assumed* model. Thus, the parameters significance test may possibly test parameters of a misspecified model. However, when the model is misspecified, the parameter estimates and their significance are irrelevant, as the model parameters do not correspond to real or population parameters. The tests of parameter significance can therefore not substitute for an overall model test.

Lacking an overall model test similar to CB-SEM, model quality in PLS is often assessed based on a set of heuristic criteria that speak to measurement and structural model properties (Gefen *et al.,* 2000). While these heuristics (reviewed below) have not been developed as a model test, they are frequently used as such. Of the 76 PLS studies we identified, almost half use the term "model test" (33 of 76) when describing their PLS analysis, showing that many IS researchers assume that the collection of these heuristics constitutes a de-facto model test.

While the extent of model misspecification in IS research can in principle not be known as we lack access to the true model, the extent of PLS use and the use of heuristics as if they constituted a model test, makes identifying model misspecification in PLS is an important issue. In this paper, we investigate to what extent the proposed PLS model quality criteria are able to identify model misspecification, both at the indicator and at the structural level. For this, we simulate different models under a variety of experimental conditions and examine the behavior of various proposed model quality criteria.

The remainder of the paper is structured as follows: The next section reviews prior simulation studies on PLS. This is followed by a review of proposed model quality criteria for PLS estimations. Then we describe the simulation study, including the models and the experimental conditions we study. This is followed by a presentation of results. The paper concludes with a discussion of the findings and recommendations for researchers.

Prior Work

Most existing studies on PLS are comparison studies that use CB-SEM results as reference. For example, a recent simulation study found that "CB-SEM clearly outperforms PLS in terms of parameter consistency and is preferable in terms of parameter accuracy as long as the sample size exceeds a certain threshold (250 observations)" (Reinarts *et al.,* 2009). However, this seeming disadvantage is made up for by the fact that "the statistical power of PLS is always larger than or equal to that of CB-SEM" (Reinarts *et al.,* 2009). The authors conclude that PLS should be preferred for small-sample research. Reinarts *et al.* (2009) also provide an overview over previous simulation studies that included CB-SEM and PLS, and find two further studies that include both techniques for a direct comparison, a recent study by Goodhue *et al.* (2006) and an early study by Areskoug (1982). Goodhue *et al.* (2006) investigate the statistical power of PLS and CB-SEM at varying sample sizes and effect sizes. Using a model with four exogenous

and a single endogenous latent variable, each measured on three reflective indicators, they find that PLS, augmented with normal-theory significance testing instead of bootstrapping, performs better than CB-SEM at sample sizes less than 200. Above that threshold, the advantage diminishes and CB-SEM is preferred due to the better accuracy of the estimates. The early study by Areskoug (1982) uses an even simpler model than Goodhue *et al.* (2006), consisting of only a single exogenous and a single endogenous latent variable. The number of indicators is varied from 4 through 32 and sample size is varied from 25 through 800. The study shows that PLS estimates with only four indicators are greatly biased and this bias is reduced as the number of indicators increases. However, the sampling variability of the estimates for small sample sizes is lower for PLS than for CB-SEM.

All three comparative studies have examined a variety of conditions, typically focusing on sample size (Areskoug, 1982; Goodhue *et al.,* 2006; Reinarts *et al.,* 2009), number of indicators (Areskoug, 1982; Reinarts *et al.,* 2009), size of effects between latent variables (Goodhue *et al.,* 2006) loadings of items on latent variables (Goodhue *et al.,* 2006; Reinarts *et al.,* 2009), distribution of observed values (Reinarts *et al.,* 2009). The latest study by Reinarts *et al.* (2009), criticizing the simple model structures used by Areskoug (1982) and Goodhue *et al.* (2006) as unrealistic, uses a model with one exogenous latent variable and five endogenous latents, containing multiple paths and mediation relationships.

A non-comparative study by Chin and Newsted (1999) reported two simulations, varying sample size, the number of latent variables, and the number of indicators per latent variable. The aim of the study was to determine the statistical significance and accuracy of the estimated item loadings, primarily for small sample sizes. Chin and Newsted (1999) conclude that PLS can produce significant estimates even for samples as small as 50.

All comparative studies (Areskoug, 1982; Goodhue *et al.,* 2006; Reinarts *et al.,* 2009) as well as that by Chin and Newsted (1999) estimate the ”true” model, from which data was generated. However, in realistic research settings, this ”true” model (”reality”) is unknown and researchers rely on the data analysis technique to identify whether their model is a good model, i.e. in close agreement with observations. A study by Cassel et al. (1999) examines the robustness of parameter estimates with respect to the skewness of the distribution of the observed variables, multi-collinearity among variables and misspecification of the structural model. They conclude that PLS estimates are quite robust to structural misspecification unless very important regressors are omitted. Their study examined the parameter bias of the misspecified model, but did not examine whether or how misspecifications can be identified in the first place. Furthermore, the study is limited to structural misspecification and does not examine measurement model misspecifications, e.g. cross-loadings. Finally, a recent study by Aguierre-Ureta and Marakas (2008) examines misspecification in the PLS context, but is limited to measurement misspecification of formative or reflective indicators. It does does not include structural misspecification or cross-loading misspecification of the measurement items.

Model Quality Criteria for PLS

There are a variety of model quality criteria for PLS, with a summary provided in Table 1. The average variance extracted (AVE) of a latent variable can be used to assess the quality of models (Chin, 1998). The AVE of each latent variable should be greater than 0.5 (Chin, 2010; Chin, 1998; Gefen and Straub, 2005; Henseler *et al.*, 2009) and its square root should be greater than the correlations of that latent variable with other latent variables (Chin, 1998; Chin, 2010; Fornell and Larcker, 1981; Gefen and Straub, 2005; Henseler *et al.*, 2009; Hulland, 1999). A more conservative heuristic is proposed by Gefen *et al.* (2000) who suggest that the AVE itself should be greater than correlations with other latent variables. Finally, Fornell and Larcker (1981) argue that the AVE of a latent variable should be larger than its r^2.

Chin (1998) suggests that the r^2 values for endogenous latent variables are a criterion for assessing the structural model. Generally, the higher the r^2 the better the predictive ability of the model (Goetz *et al.,* 2010). However, no guidelines exist for what value of r^2 is acceptable, although the statistical significance of r^2 can be determined using an F-test (Fornell and Larcker, 1981).

Another measure of model quality, based on factor-analytic considerations, is the construct reliability, assessed using either Cronbach's α (Goetz *et al.,* 2010; Gefen *et al.,* 2000; Straub *et al.,* 2004) or the composite reliability or internal consistency metric ρ (Goetz *et al.,* 2010; Chin, 1998; Chin and Gopal, 1995; Fornell and Larcker, 1981, Henseler *et al.*, 2009). As a guideline, construct reliability should exceed 0.7 (Hulland, 1999; Gefen *et al.,* 2000; Straub *et al.,* 2004) or 0.6 (Goetz *et al.,* 2010, *Henseler et al.*, 2009). Other criteria based on factor-analytic

considerations suggest that the loadings of item on its latent variable should exceed .707, so that the latent variable accounts for more than 50% of the item variance (Goetz *et al.,* 2010; Hulland, 1999; Gefen *et al.,* 2000; Chin and Gopal, 1995; *Henseler et al.*, 2009; Straub *et al.,* 2004). On the other hand, the loadings of items on latent variables they are not indicators of should be comparatively low (Chin, 2010; Goetz *et al.,* 2010; Hulland, 1999; Gefen *et al.,* 2000; Gefen and Straub, 2005; Henseler *et al.*, 2009; Straub *et al.,* 2004). Gefen and Straub (2005) suggest a relative criterion whereby item loadings should be higher than cross-loadings by at least 0.1. For principal components analysis, Straub *et al.* (2004) suggest that cross-loadings should be lower than 0.4.

Table 1: Criteria for assessing quality of models using PLS

Criterion		**Recommendation**
1	(ROOT) AVE	> latent correlations
2	(ROOT) AVE	> 0.5
3	(ROOT) AVE	> r^2
4	r^2	F-test significant, larger is better
5	Reliability (α)	> 0.7
6	Internal Consistency (ρ)	> 0.7
7	Item loadings	> 0.7
8	Item loadings	Greater than cross-loadings (by at least 0.1)
9	Cross loadings	< 0.4
10	GoF (various)	

Finally, recognizing the need for an overall measure of model quality, goodness-of-fit (GoF) metrics (absolute, relative, measurement and structural GoF) have been proposed that takes into account the PLS optimization objective and is a measure of the "achievable fit" (Tenenhaus *et al.,* 2004; Esposito Vinzi *et al.,* 2010). Chin (2010) and Esposito-Vinzi et al. (2010) propose that a relative GoF of .90 or higher is "suggestive of a good model".

Study Design

The recent study by Reinarts *et al.* (2009) provides an overview of design factors and outcome variables of PLS (and CB-SEM) simulation studies. They note that previous simulation studies using PLS have primarily focused on the accuracy of parameter estimates given a properly specified model. Other issues examined are statistical power and convergence of solutions. Their study also focuses on parameter accuracy and statistical power. They identify four design factors from past studies and apply them in their study: (1) Sample size, (2) Number of indicators, (3) Indicator distribution, and (4) Indicator loadings.

Table 2: Study Design Factors

Design factor		**Levels**
s	Sample size	100, 250, 750
i	Number of indicators	3, 5, 7
l	Indicator loadings (non-std.)	.75, 1, 1.25
b	Effect sizes (β,γ)	.25, .75
—	Sampling distribution	Normal
—	Response type	Continuous
—	Missing values	None

Besides these, previous simulation studies have varied the effect sizes between latent variables (Goodhue *et al.,*

2006; Curran *et al.,* 2002), the distribution of samples, their skewness and/or kurtosis (Flora and Curran, 2004; Lei and Lomax, 2005; Savelei, 2008; Gold *et al.,* 2003; Cassel et al., 1999), the response type (categorical/continuous) (Flora and Curran, 2004), the CB-SEM optimization criteria (Olsson *et al.* 2004; Lei and Lomax, 2005; Fan *et al.,* 1999), and the proportion of missing values (Davey *et al.,* 2005; Savalei and Bentler, 2005; Savelei, 2008; Gold *et al.,* 2003; Dolan *et al.,* 2005). For the present study, we are not interested in the effects of missing values, different optimization criteria, or distribution of variables. The remaining design factors will be varied, as shown in Table 2. In the interest of keeping computational and storage requirements manageable, these factors will be varied on a limited number of levels only. We use only reflective indicators in this study. Error variance of all indicators is set to .1 and endogenous latents also have an error (disturbance) variance of 0.1.

Models

We wish to investigate typical models, ranging in complexity from simple to complex. However, there are no metrics of model complexity to help with model selection (Fan *et al.,* 1999). Much of the simulation literature has focused on CB-SEM studies and many of those studies have been limited to CFA-type models. More complex models are those by Reinarts *et al.* (2009), who examine a six factor model with one exogenous and five endogenous factors. Areskoug (1982) examines a simple two factor model. In the CB-SEM literature, models that go beyond a CFA-style model are a three factor chain (Curran *et al.,* 2002) a three factor chain with direct effect from exogenous to final endogenous factor (Lei and Lomax, 2005), a four factor model with two exogenous and two endogenous factors (Fan *et al.,* 1999), and a four factor model with one exogenous and three endogenous factors (Gold *et al.,* 2003). Based on this literature, we examine the models shown in Figures 1 through 3, which we believe are realistic models and show different degrees of complexity.

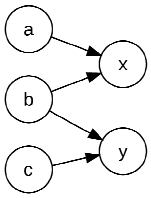

Figure 1: Model 1

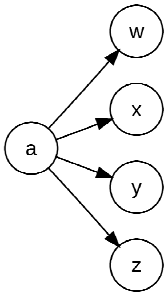

Figure 2: Model 2

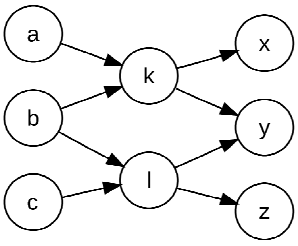

Figure 3: Model 3

Model Misspecification

Model misspecification is difficult to quantify, as there are numerous ways in which a model may be misspecified and there is no metric for the degree of misspecification (Fan *et al.,* 1999). There has been no work on the effect of model misspecification in PLS and, given that many CB-SEM studies focus on CFA-style models, much of the research on misspecification has focused on item cross-loadings. Even the simulation studies that examine relatively complex models only specify cross-loadings (or their absence) as kinds of misspecifications. Based on this literature, we examine the PLS results for the correct model (condition 0), and the following misspecification conditions:

- Model 1
  - True model contains one cross-loading from *A* on *C* and from *X* on *Y* (condition M1XL1)
  - True model contains two cross-loadings from *A* on *C* and from *X* on *Y* (condition M1XL2)
  - True model contains two cross-loadings from *A* on *X* and from *C* on *Y* (condition M1XL3)
  - True model contains latent path from *A* to *Y* (condition M1L1)
  - True model contains latent path from *A* to *B* (condition M1L2)

    - True model contains latent path from *X* to *Y* (condition M1L3)
    - Combination of L1 — L3 (condition M1L4)
    - True model reverses paths between *B* and *X* and *B* and *Y* (condition M1R1)
- Model 2
    - True model contains one cross-loading from *W* on *X* and from *Y* on *Z* (condition M2XL1)
    - True model contains one cross-loading from *A* on *X* and from *A* on *Z* (condition M2XL2)
    - True model contains latent path from *W* to *X* and *Y* to *Z* (condition M2L1)
    - Estimated model reverses path between *A* and *W* (condition M2R1)
    - Estimated model reverses paths between *A* and *W* and *A* and *X* (condition M2R2)
- Model 3
    - True model contains one cross-loading from *A* on *C* and from *X* on *Y* (condition M3XL1)
    - True model contains two cross-loadings from *A* on *C* and from *X* on *Y* (condition M3XL2)
    - True model contains two cross-loadings from *A* on *X* and from *C* on *Z* (condition M3XL3)
    - True model contains latent paths from *A* to *L* and *C* to *K* (condition M3L1)
    - True model contains latent paths from *A* to *Z* and *C* to *X* (condition M3L2)
    - True model contains latent paths from *K* to *Z* and *L* to *X*, estimated model does not (condition M3L3)
    - True model reverses paths from B to K, B to L, K to Y, and I to Y (condition M3R1)

We estimated each simulation condition using PLS, noting the evaluation criteria in Table 1. We generated 100 samples from the true model for each experimental condition. We used the R statistical system (version 2.10), with the `plspm` (version 0.1-4) package. For each evaluation criterion, we compared the mean of the 100 samples for the true model to the mean of the 100 samples for the misspecified model.

Results

Because of the large number of experimental conditions (3×3×3×2=54 parameter variations for each condition/model, and 8+6+8=21 conditions/models for a total of 54×21=1134 conditions), we do not provide complete results in tabular form. Instead, the following subsections provide a brief discussion of the behavior of the proposed model evaluation criteria for the different models and misspecification conditions. We conducted ANOVA analyses to identify the effects of the different experimental variables on the model quality criteria. Again, for lack of space, in Table 3 below we only present one such result for the relative goodness-of-fit criterion. Instead, we consider graphical presentations, as in Figures 4–23 more informative.

AVE relative to latent correlations (Criterion 1)

For this criterion, we examined the percentage of latent variables for which the AVE was greater than all the correlations of that latent with other latent variables. Thus, we expect this criterion to decrease for misspecified models. For the true M1, the criterion was 1 for all conditions, but for the true M2 and M3, the criterion was less than 1, as low as 0.34 in some conditions for the true M2 and as low as 0.93 for some conditions for the true M3. This makes it difficult to define a cutoff that identifies a misspecified model.

There was no or only a minor (< 5% change) effect for M1XL1, M1XL2, M1L1, M1L2, M1L3, M1L4, M3XL1, M3XL2, M3L2, M3L3, M3R1 conditions. For M1XL3, there was a minor decrease (~8%) only for b=0.25 and l=0.75. There was minor decrease (up to ~16%) for M1R1 when b=0.25 and l=0.75, decreasing with i and s. For M2XL1 and M2XL2, there was a decrease up to 70% for b=0.75, l=0.75 and increasing with i and s. For M2L1, there was a decrease up to 28% only for b=0.75 and with an interaction effect between i, s and l, as shown in Figure 4 below. A more complex effect was evident for M2R1, where the criterion *increased* for some conditions, and decreased for other conditions, shown in Figure 5 below. For M2R2, the criterion *increased* by up to 150% when l=0.75, b=0.75 and i=7. There were mixed but minor effects also for M3XL3, shown in Figure 6. A decrease up to 16% occurred for M3L1, only for b=0.75 and this effect weakened as l increased.

Given the range of this criterion for the true models, and the fact that it improved considerably for some misspecifications, make this criterion, by itself, appear ill-suited for identifying model misspecifications.

AVE absolute (Criterion 2)

For this criterion, we examined the percentage of latent variables for which the AVE was greater than 0.5. We expect this criterion to decrease for misspecified models. Again, for the true M1, this criterion was 1 for all conditions, but for the true M2 it dropped to as low as .93 for some conditions while it was as low as .75 for some conditions of the true M3.

There were no or only minor (< 5% change) effects for M1XL1, M1XL2, M1XL3, M2XL1, M2XL2, M2L1, M3XL1, M3XL2, and M3R1. We observed a decrease by up to 50% for M1L1, M1L2, M1L3, M1L4, M1R1 when b=0.75. This effect decreased with increasing s and decreasing i. Figure 7 shows the effect for M1L2. There was a decrease for M2R1 (up to ~60%) and M2R2 (up to ~40%) , when b=0.25 and l=0.75. The criterion showed complex interaction effects for M3XL3 (Figure 8). M3L1, M3L2, M3L3 showed an *increase* in the criterion up to 30% for b=0.25 and l=0.75.

Similar to the previous criterion 1, the range of the AVE for true models and the fact that it increased for some conditions would seem to make this criterion unsuitable for determining whether a model is misspecified.

AVE relative to r^2 (Criterion 3)

For this criterion, we examined the percentage of latent variables for which the AVE was greater than their r^2 value. Again, we expect a decrease in this criterion for misspecified models. For the true M1, M2 and M3, this criterion was 1 except when b=0.75, l=0.75, and i=7, when it dropped to approx. 0.75.

For M1XL1 we saw a decrease by up to 50% when b=0.75, which diminished with increasing l. Similarly, for M1XL2 there was a decrease by up to 50% for b=0.75. The criterion decreased by up to 70% for M1XL3, but only for l=0.75, with a stronger effect for b=0.75. There was an *increase* in the criterion for M1L1, M1L2, M1L3, M1L4, M1R1, M2XL1, M2XL2, M2XL3, M2R1, and M2R2 when b=0.75 and l=0.75. Figure 9 shows the effect for M1R1. This criterion decreased for M3XL1, M3XL2, and M3XL3 when b=0.75 and l=0.75. For M3L1 the criterion showed complex interactions, increasing for some conditions and decreasing for others. The criterion *increased* for M3L2, M3L3 and M3R1 when l=0.75 and i=7, with stronger effects for b=0.75.

Criterion 3 behaved more reliably than criteria 1 and 2, with increases limited to a single condition and decreases for misspecifications otherwise. However, this criterion was insensitive to many of the misspecifications.

R^2 (Criterion 4)

For this criterion, we began by examining the percentage of endogenous latent variables whose r^2 was statistically significantly different from 0. This was 1 for all conditions of the true and misspecified M1. It was also 1 for the true M2. For M2R1 and M2R2 the percentage of significant r^2 dropped by 3% to 25% when b=0.25 and s=100 or s=250. For the true M3, the percentage of endogenous variables with significant r^2 was 1 except for s=100 and b=0.25 when it dropped to approx. 0.95. However, there was no difference to this percentage for the misspecified conditions of M3. Hence, we believe the significance of r^2 is not a sufficient criterion to identify misspecifications.

We next compared the true M1, M2 and M3 with the misspecified conditions with respect to the average r^2 of all endogenous variables. For M1XL1, M1XL2, M1XL3, M2XL1, M2XL2, M2XL3, the average r^2 *increased* by up to 35% for the misspecifications, with a stronger effect for b=0.25. For M2L1, the r^2 decreased up to 4% when b=0.75 and increased by up to 20% for b=0.25. For M2R1, the average r^2 decreased by up to 60% for b=0.25 and showed no change for b=0.75. For M2R1, the average r^2 decreased by approx 45% for b=0.25 and by approx 20% for b=0.75. The misspecifications M3XL1 and M3XL2 showed little change in average r^2 (< 3%), but the average r^2 *increased* by up to 35% for M3XL3 when b=0.25. M3L1, M3L2, M3L3 showed a decrease in average r^2 with a stronger effect when b=0.75 and M3R1 showed an average decrease of r^2 of about 10% with complex interactions of the experimental conditions, shown in Figure 10.

The test for significance appears to be meaningless as for most realistic models, the r^2 of the endogenous latent variables is almost certainly significant. Using the average r^2 value to identify misspecifications also appears to be problematic, as there is no cutoff point, the average r^2 varies considerably among true models, and the average r^2 improves for a number of misspecifications. Moreover, the r^2 , while usually considered a criterion related to the structural or inner models, is sensitive also to the misspecifications of the measurement or outer model.

Reliability (α) (Criterion 5)

For this criterion, we examined the percentage of latent variables whose reliability was > 0.7. This criterion was 1 or 100% for the true M1. There was no effect for any of the M1 misspecifications; all latent variables had a reliability > 0.7.

For the true M2, this criterion was 1 except for *b*=0.25, *i*=3, *l*=0.75 and *s*=100 where it was approx. 0.8. For these conditions, the M2XL1, M2XL2, and M2L1 misspecifications showed a minor *increase* (up to approx. 8%) in this criterion. For the M2R1 and M2R2 misspecifications, there was a decrease in this criterion by up to 60%, limited to *b*=0.25 and *i*=3. For the true M3 when *i*=3, *l*=0.75 and *b*=0.25, the percentage of latent variables with reliability > 0.7 is only approx .70. Only in those conditions is there a difference in the misspecifications M3XL2, M3XL3, M3L2, and M3L3 where this percentage *increases*.

We also examined the decreases in average reliability of all latent variables for the misspecification conditions. There were decreases of up to 2.5% for M1XL1, M1XL2, M1XL3, M2XL1, M2XL2 with the effect strengthening as *i* increases. We observed *increases* of up 10% for M1L1, M1L2, M1L3, M1L4, and M1R1, stronger for *b*=0.25 and decreasing with loadings *l*. For M2L1, the average reliability increased by up to 2%, a stronger effect for lower *l*. For M2R1, the average reliability decreased by up 7%, stronger effects for *b*=0.25 and decreasing with higher *l* (Figure 11). The changes to average reliability were within 2% for all M3 misspecifications.

This is another criterion that responds in different directions depending on experimental condition and misspecification. Moreover, while reliability is a measurement issue, this criterion is also sensitive to some structural misspecifications. Finally, the decrease of this criterion for actual measurement misspecification was small, making it possible that this could be overlooked it practice.

Internal Consistency (ρ) (Criterion 6)

This criterion is similar to criterion 5 above. We examine the percentage of latent variables whose internal consistency was > 0.7. This criterion was 1 or 100% for the true M1 but decreased by up to 40% for M1L2, M1L4, and M4R1. The effect is stronger for *b*=0.75 and lower *l*. This criterion was 1 for the true M2 and M3 and for all misspecifications of M2 and M3.

We also examined the decreases in average internal consistency of all latent variables for the misspecification conditions. For the true models M1, M2, and M3 the average internal consistency was between 0.97 and 0.99. There were no differences for M1XL1, M1XL2, and M1XL3. However, the average consistency decreased by up to 70% for the M1L1, M1L2, M1L3, M1L4, and M1R1 misspecifications. The effect was stronger for *b*=0.25. There was no effect for M2XL1, M2XL2 and M2L1 but a decrease of up to 12% for M2R1, again with a stronger effect for *b*=0.25 and also stronger for lower *l* (Figure 12). There were no effects for any of the misspecifications of M3.

Because Cronbach's α(criterion 5) is a lower bound of reliability, the internal consistency is always higher. However, it appears that it is so high that it becomes a meaningless criterion as all latent variables in all misspecified models had an internal consistency > 0.7, sufficient to satisfy this criterion.

Absolute item loadings (Criterion 7)

For this criterion we determined the percentage of all loadings in a model above a 0.7 threshold. For the correct model, this criterion is 1 or 100%. For the true M1, this criterion was 1 for all experimental conditions, for the true M2 it was approx. 0.97 (when *b*=0.25, *l*=0.75, *i*=3), 0.89 (when *b*=0.25, *l*=0.75, *i*=5) and 0.81 (when *b*=0.25, *l*=0.75, *i*=7). For the true M3 it was approx. 0.9 when *b*=0.25, and *l*=0.75.

The criterion did not change for the M1XL1, M1XL2, M1XL3, M1L1, M1L3, M2XL1, M3XL3, M3L1, M3L2, and M3L3 misspecifications. This criterion was affected strongly by the M1L2 and M1L4 misspecification conditions when *b*=0.75 where the criterion dropped by up to 1/3 to approx. 0.67 (Figure 13). The effect was stronger with increasing *l* and *s*. This criterion also strongly picked up the M1R1 misspecification, again only for *b*=0.75, where the percentage of loadings >.7 dropped by approx. 2/3, again a stronger effect for increasing *l*, and *s*. However, M1L2, M1L4, and M1R1 are structural misspecifications, yet this criterion is primarily concerned with measurement properties and might thus yield a false indication of the misspecification.

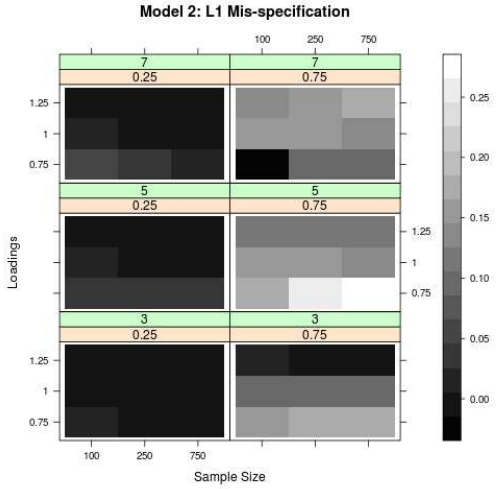

Figure 4: Criterion 1 percent decrease for M2L1

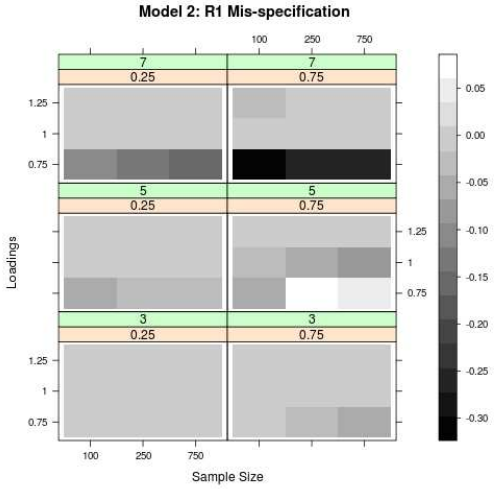

Figure 5: Criterion 1 percent decrease for M2R1

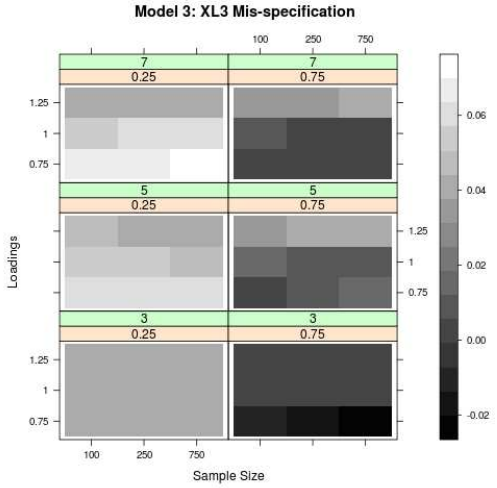

Figure 6: Criterion 1 percent decrease for M3XL3

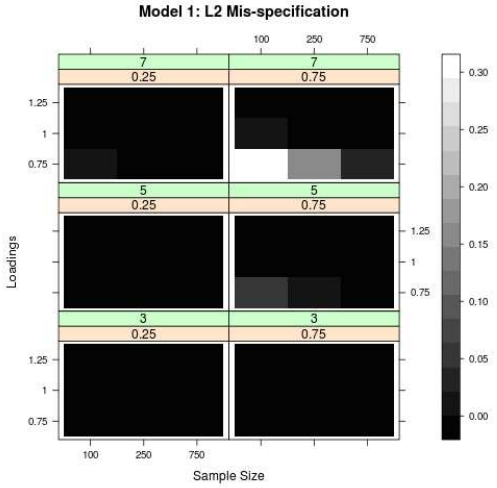

Figure 7: Criterion 2 percent decrease for M1L2

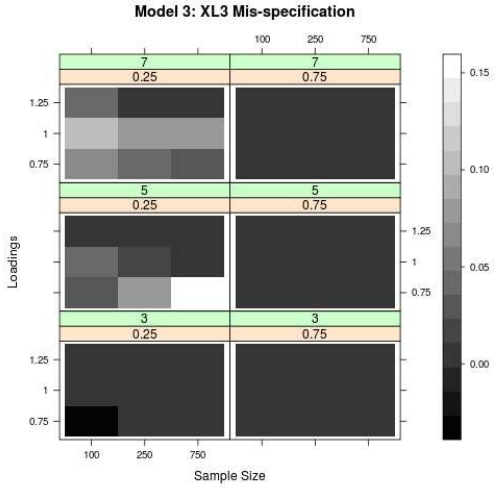

Figure 8: Criterion 2 percent decrease for M3XL3

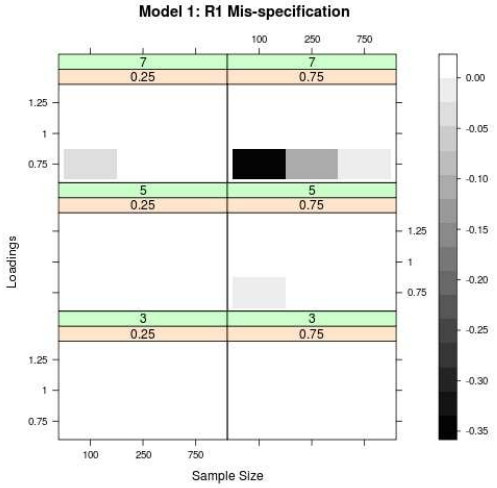

Figure 9: Criterion 3 percent decrease for M1R1

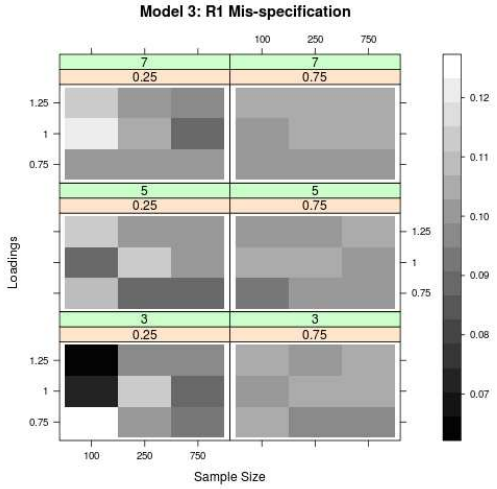

Figure 10: Criterion 4 percent decrease for M3R1

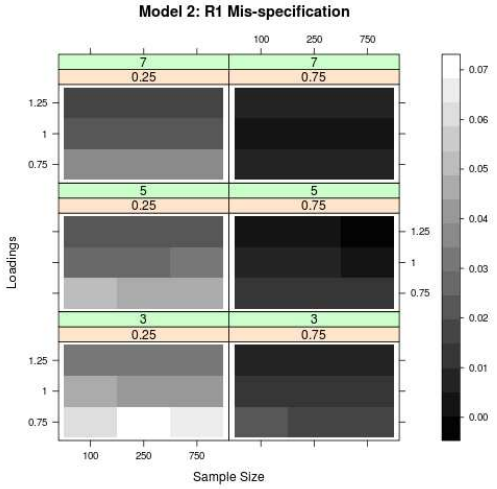

Figure 11: Criterion 5 percent decrease for M2R1

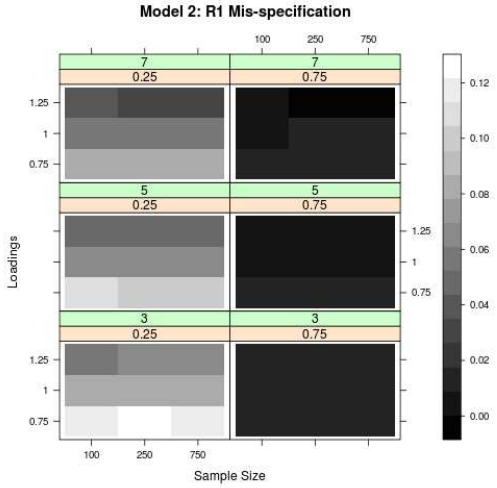

Figure 12: Criterion 6 percent decrease for M2R1

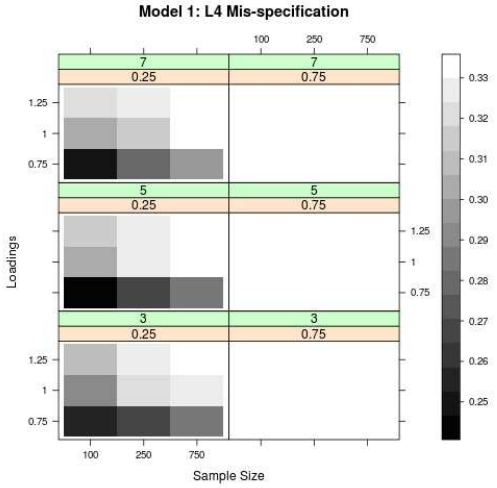

Figure 13: Criterion 7 percent decrease for M1L4

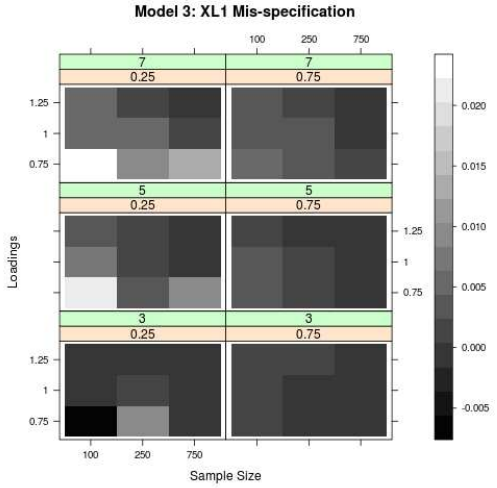

Figure 14: Criterion 7 percent decrease for M3XL1

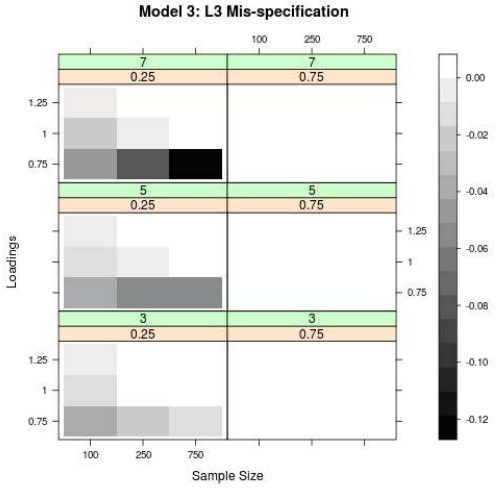

Figure 15: Criterion 7 percent decrease for M3L3

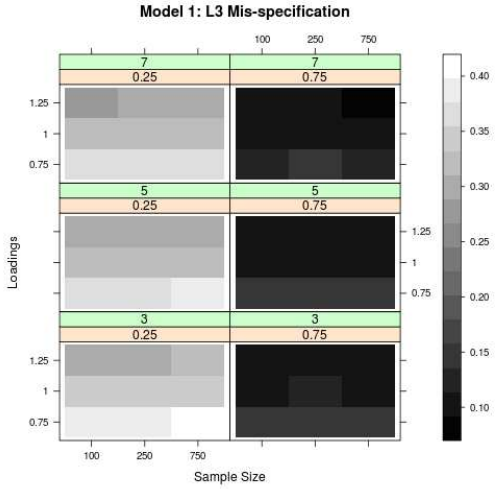

Figure 16: Criterion 8 percent decrease for M1L3

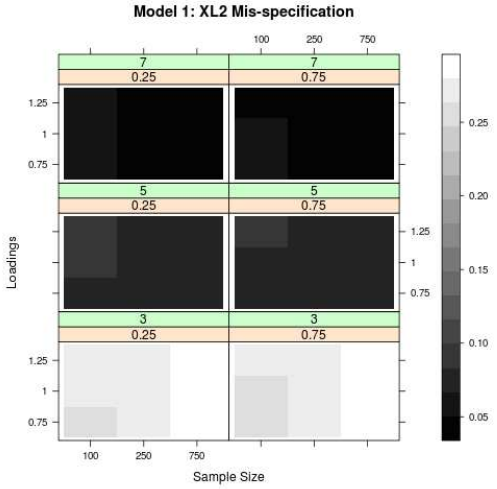

Figure 17: Criterion 9 percent decrease for M1XL2

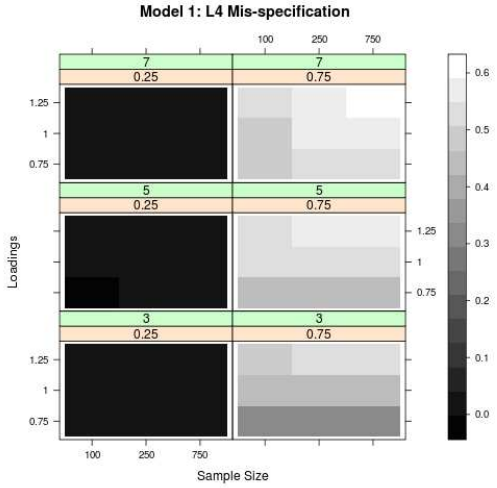

Figure 18: Criterion 9 percent decrease for M1L4

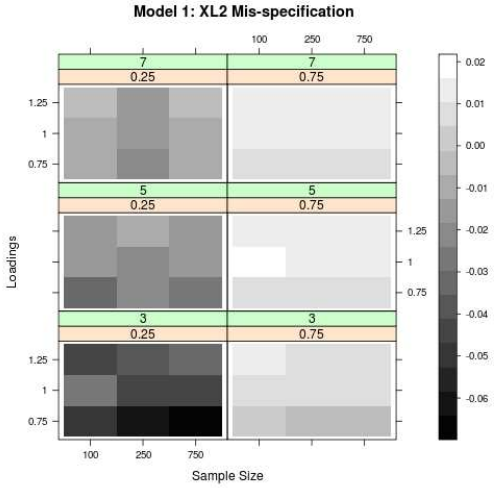

Figure 19: Criterion 10a percent decrease for M1XL2

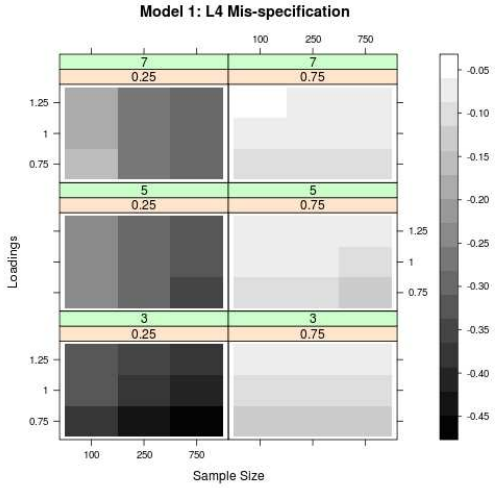

Figure 20: Criterion 10a percent decrease for M1L4

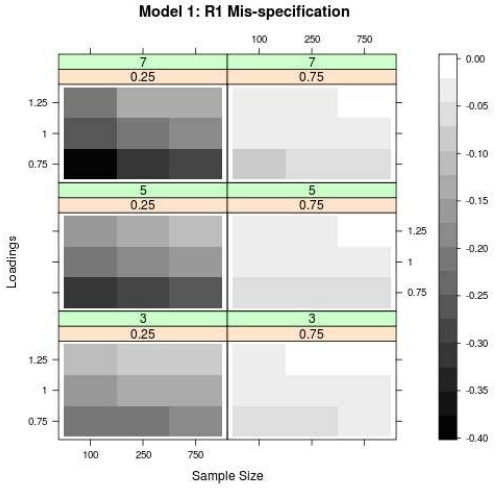

Figure 21: Criterion 10b percent decrease for M1R1

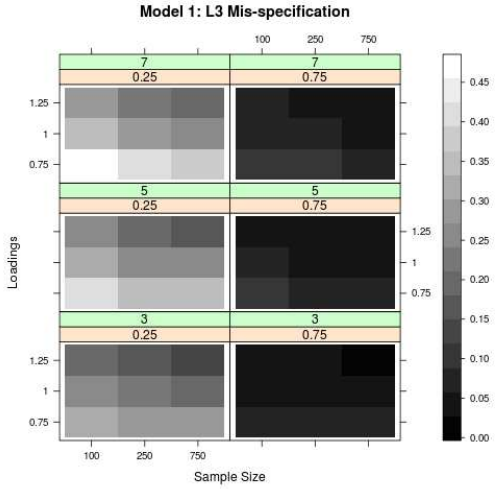

Figure 22: Criterion 10c percent decrease for M1L3

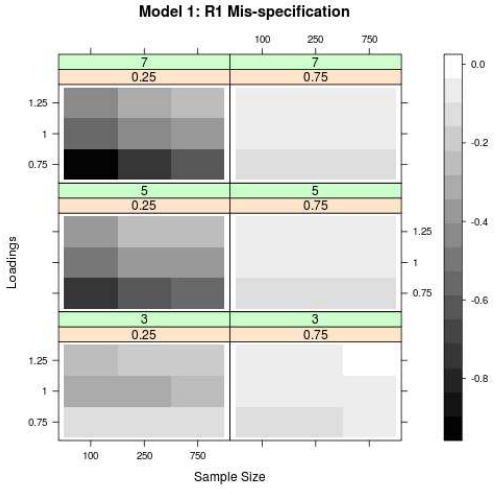

Figure 23: Criterion 10d percent decrease for M1R1

The criterion showed a decrease by up to 5% for M3XL2 when *b*=0.25, the effect decreasing with *l*. For M2L1, there was an increase by up to 7% when *b*=0.25, an effect decreasing with *l* and *s*. There was a strong decrease in this criterion for M2R1 (up to 45%) and M2R2 (up to 25%) when *b*=0.25. The effect decreased with increasing *l*. Again, a criterion intended to assess cross-loadings appeared primarily sensitive to structural misspecifications.

This criterion was 1 or 100% for the true M3, except for *l*=0.75 and *b*=0.25 in which case it was approx. 0.97 (*i*=3), 0.90 (*i*=5) or 0.82 (*i*=7). For M3, this criterion picks up the M3XL3 misspecification, but only for *b*=0.25 and with interaction effects with *l*, *s*, and *i*. Here, the mean criterion values for the misspecified model are about 10% to 20% lower than for the true M3. There were minor effects (less than 5% decrease) also for M3XL1 (Figure 14) and M3XL2. There is an increase in this criterion by up to 5% for M3L1 when b=0.25 and by up to 15% for M3L2 and M3L3 (Figure 15), with stronger effects for low *s* and *l*.

Given that factor loadings are a measurement issue, it is surprising that this criterion was very sensitive to structural misspecification, e.g. for M1L2, M1L4, M3L2, and M3L3, but not to many measurement misspecifications, e.g. M1XL1, M1XL2, etc. Thus, this criterion appears to not only be unreliable in identifying misspecifications but also possibly misleading in indicating the cause of the misspecification.

Item loadings relative to cross-loadings (Criterion 8)

For this criterion we determined the percentage of cross loadings that are smaller than the intended factor loadings by at least 0.1. We examined cross-loadings only for the indicators of exogenous latent variables. Thus, this criterion is not applicable to M2. For the correct model, this percentage should be 1 or 100% and we found this for the true M1 but not for the true M3 where, for *b*=0.75, the criterion was approx. 0.97.

We found weak effects (less than 2% change) for the M1XL1, M1XL2, and M1XL3 misspecifications. There were strong decreases in this criterion for M1L1, M1L2, M1L3, M1L4 and M1R1. The criterion dropped by about 40% to 50% for when *b*=0.25 and by about 5% - 20% when *b*=0.75 with interaction effects of *s* and *l* (Figure 16). For M3, this criterion is affected in very minor ways (less than 2% change) by the M3XL1, M3XL2, M3XL3, and M3R1 misspecifications with slightly larger effects in the large structural effect condition (*b*=0.75). The criterion showed strong decreases by up to 25% for M3L1 and M3L3 misspecifications when *b*=0.75. For M3L2 there was an *increase* in this criterion when *b*=0.75, the criterion was 1 in this misspecification but not in the true M3.

Similar to criterion 7, while the criterion is intended to identify measurement misspecifications, it appears to be most sensitive to structural misspecifications instead.

Cross loadings (Criterion 9)

For this criterion we determined the percentage of cross loadings smaller than 0.4. We examined cross-loadings only for the indicators of exogenous latent variables. Thus, this criterion is not applicable to M2. For the true model, this criterion should be 1 or 100%. We found this for the true M1 and the true M3.

This criterion was not affected by the M1XL3, M1L1, and M1L3 misspecifications. For M1XL1 there was a small decrease (up to 7%) , which was stronger for low *i*. Similarly, there as was decrease by up to 27% for M1XL2, stronger again for low *i* and some interaction effect of *s* and *l* (Figure 17). The criterion decreased by up to 30% for M1L2, up to 60% for M1L4, and up to 20% for M1R1 when for *b*=0.75 and a stronger effect with increasing *s* and *l (*Figure 18). For M3, this criterion was not affected by the M3XL3, M3L1, M3L2, and M3L3 misspecifications. There was a decrease in this criterion for M3XL1 (up to 6%) and M3XL2 (up to 25%), with a stronger effect for lower *i*. For M3R1, the criterion decreased by up to 25%, only when *b*=0.75 and with interaction effects among *i*, *l*, and *s*.

Similar to the two previous criteria, this is a criterion primarily intended to identify cross-loadings and measurement model problems, but is strongly affected by some structural misspecifications.

Absolute goodness of fit (AgoF) (Criterion 10a)

This criterion is the absolute goodness of fit (AGoF) measure (Tenenhaus *et al.,* 2004). There is no standard for correct models. We found the AGoF to range from about 0.45 to 0.9 for our true models. Such a variation makes the use of this metric problematic for assessment of single models, but it may be useful for model comparison. The AGoF fit for the true M1, M2 and M3 was approx. 0.45 to 0.65 for *b*=0.25 and approx. 0.7 to 0.9 for *b*=0.75.

The AGoF criterion showed minor changes for M1XL1 and M1XL2, positive or negative depending on *b*, *i*, *l*, and *s* (Figure 19). There was an *increase* by up to 25% for M1XL3 when *b*=0.25, with a stronger effect for lower *l*. There were also increases by up to 45% for M1L1, M1L2, M1L3, M1L4, and M1R1, which again depended on a combination of *b*, *i*, *l* and *s*, and were stronger for *b*=0.25 (Figure 20).

This criterion *increased* by up to 10% for M2XL1, M2XL2, and M2L1 with a larger effect for *b*=0.25 and interaction effects for *i*, *l*, and *s*. The AGoF decreased for M2R1 (up to 40%) and M2R2 (up to 25%) with stronger effects for *b*=0.25.

The AGoF was not affected by M3XL1 or M3XL2 but showed an *increase* by up to 15% for M3XL3 when *b*=0.25, with a stronger effect for lower *l*. There were minor changes (less than 5%), positive as well as negative, for M3L1, M3L2, M3L3, and M3R1, depending on the combination of *b*, *i*, *l*, and *s*.

Overall, the AGoF is affected in different directions, but mainly positive, by different misspecifications; it is not affected by all misspecifications; and any effect can also differ with the strength of structural relationships and the number of indicators. Hence, the AGoF does not appear to be a reliable indicator of misspecifications.

Relative Goodness of Fit (RGoF) (Criterion 10b)

This criterion is the relative goodness of fit (RGoF) measure. As for the AGoF, there is no standard for correct models. For the correct models, we found the RGoF value to range from 0.60 to 0.98. Again, this lack of standard makes an absolute model assessment problematic.

There were minor changes to the RGoF for M1XL1 and M1XL2 (less than 3%). The RGoF decreased for M1XL3 when *b*=0.25, with a stronger effect for increasing *s* and *l*. For M1L1, M1L2, M1L3, M1L4, and M1R1, the RGoF *increased* by up to 40%. Here, the effect was stronger for *b*=0.25 and strengthening with decreasing *s* and *l* (Figure 21). There was a minor decrease for M2XL1 and M2XL2 (up to 4%), with stronger effects for higher *l* and *s*. M2L1 showed a small increase (up to 4%) with stronger effects for lower *s* and *l*. M2R1 showed a decrease for *b*=0.25, with a stronger effect for lower *s*. There was no effect for M3XL1 and M3XL2. The RGoF changed by up to 8% for M2XL3 and this change was positive or negative depending on the combination of *b*, *i*, *l*, and *s*. There were only minor effects (less than 5% change) for M3L1, M3L2, M3L3, and M3R1. We conducted an ANOVA to identify the effects of the different experimental variables on RGoF. Table 3 shows the main and 2-way interaction effects (Cohen's f) on RGoF. Intercepts and higher-order interaction effects exist but are omitted for space reasons.

In summary, similar to the AGof, the RGoF does not appear to be a reliable indicator of model misspecifications, as it may show no or very minor differences only for some misspecifications, and it shows a significant increase for other misspecifications.

Table 3: Effect sizes for effects on RGoF (number of + or – indicates f and direction of effect)

Model	Cond.	Sample size	indicators	loadings	beta	Sample size x indicators	Sample size x loadings	Indicators x loadings	Sample size x beta	Indicators x beta	Loadings x beta
M1	XL1	- -	- -	+ +	+	+	+	+	+	+ +	
M1	XL2	- -	+ +	+	+			-	+	-	
M1	XL3	+	+ +	+ + + + + + +	+			-	-	-	
M1	L1	- -	- - -	+ + + + + +	+		+	+	+	+ +	
M1	L2	- -	- - -	+ + + + + +	+		+	+	+	+ +	
M1	L3	- -	- - -	+ + + + + +	+		+	+	+	+ +	
M1	L4	- -	- - -	+ + + + + +	+		+	+	+	+ +	
M1	R1	- -	- - -	+ + + + + +	+		+	+	+	+ +	
M2	XL1	+	+ +	+ + + +	+			-	-	-	
M2	XL2	+ +	+ +	+	+			+	-	-	
M2	L1	- -	- - -		-		+	+	-	-	
M2	R1	-		- - -	-			+ +	+	+	
M2	R2	-	-		+		+	+	+	+	
M3	XL1		+	-		+		+			
M3	XL2	-	-					+	+	+	+
M3	XL3	+	+	+			+	-	-	-	
M3	L1		- -	- - -	+		+	+	+		
M3	L2	-	-		-		+	+	+		+
M3	L3	+	+	+ +				-	-	-	
M3	R1	-	- -	- - - - - -	+		+	+	+	+ +	+

Structural Goodness of Fit (Criterion 10c)

The two overall goodness-of-fit indices (AGoF and RGoF) are composed of individual goodness-of-fit metrics for the structural (inner) and the measurement (outer) model in PLS. For the true M1, M2 and M3 the structural GoF (SGoF) was 1 for all conditions. The SGoF showed no change for M1XL1, M1XL2, and M1XL3. It decreased by up to 60% for M1L1, M1L2, M1L3, M1L4 and M1R1. In all cases, there was a stronger effect when b=0.25 and for lower s and l (Figure 22). There was no change for M2XL1, M2XL2, M2L1, M2R2 and only a minor decrease (less than 4%) for M2R1 when b=0.25, with a stronger effect for lower s and l. The SGoF showed no change larger than a 1.5% change for any M3 misspecification.

The SGoF criterion does appear to be sensitive primarily to structural misspecifications, as intended. However, the fact that it did not pick up any of the M3 misspecifications means that it is not a very reliable criterion.

Measurement Goodness of Fit (Criterion 10d)

The measurement goodness-of-fit (MGoF) is another component of the overall goodness-of-fit measures. The MGoF of our true models ranged from 0.44 to 0.96. This range makes this criterion problematic for absolute model evaluations.

There were minor decreases (up to 5%) for M1XL1 and M1XL2, varying with the combination of *b, i, l,* and *s*. There was a strong effect (decrease up to 35%) for M1XL3, with a stronger effect for *b*=0.25 and strengthening with increasing *s* and *l*. The MGoF *increased* by up to 80% for M1L1, M1L2, M1L3, M1L4 and M1R1, with a stronger effect for *b*=0.25 and strengthening with decreasing *s* and *l* (Figure 23). The MGoF criterion decreased by up to 8% for M2XL1 with stronger effects for *b*=0.25 and strengthening with *s* and *l*. There were mixed effects (positive and negative changes) for M2XL2 and M2L1, up to 6% in either direction. Generally, the MGoF *increased* with higher *l* and *s* and decreased with lower *l* and *s*. For M2R1 there was a decrease by up to 35%, only when *b*=0.25 and with a stronger effect for lower *s*. There were mixed effects (up to 20% increase or decrease) for M2R2 with stronger effects for *b*=0.25 and an increase for lower *s* and higher *l* and a decrease for higher *s* and lower *l*. The MGoF showed only minor changes (less than 2%) for M3XL1 and M3XL2. Again, there were mixed effects for M3XL3, only for b=0.25. In this case, the MGoF decreased when *l* and *s* are high and increased when *l* and *s* are low. Similarly complex behaviors were seen for M3L1, M3L2, M3L3 and M3R1, with the MGoF increasing or decreasing depending on the combination of *b*, *i*, *l* and *s*.

The complex behavior of this criterion and the fact that it improves for some misspecification conditions appears to make it unsuitable to reliably detect model misspecification.

Advanced Analyses

Beyond the descriptive and ANOVA approach reported on above, we analyzed our data using a number of alternative techniques. A classification tree approach was performed to construct decision trees for identifying classes of data (Ripley, 1996). Figure 24 shows a decision tree for Model 1, with a misclassification error rate of .14. Besides the high error rate, tree models and their decision rules are model specific, i.e. the decision variables and cutoff points differ between models, so that no clear guidance can be given to researchers. While we believe our models are comparable in complexity to realistic models in IS research, it is unlikely that a real model will be exactly like one of the three we specified. Moreover, the tree mirrors the problems found in the analysis above. For example, while Criterion 8 (item loadings relative to cross loadings) is used to separate structural from measurement (cross-loading) misspecifications, one would expect cross-loadings to be worse for measurement misspecifications. Instead, by tree in Figure 24, cross-loadings are worse for structural misspecifications. Criteria 7 and 9, which are the other two criteria used in this tree, are both cross-loading criteria. Yet, they can apparently be used to identify different structural misspecification.

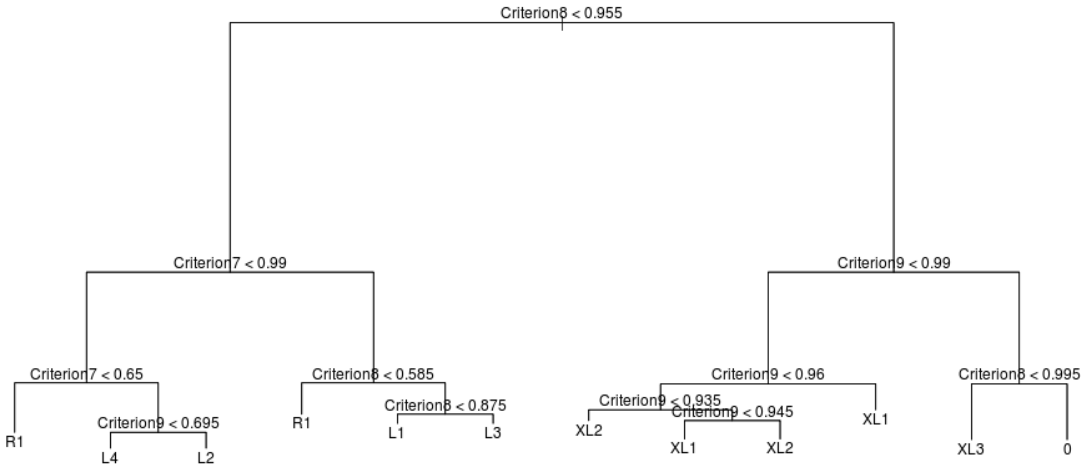

Figure 24: Decision tree for identifying misspecifications in Model 1

Another form of classification analysis is linear discriminant analysis (LDA) (Venables and Ripley, 2002). The results of the LDA are similar to those of the classification tree analysis. For example, the first linear discriminant for Model 1 is most strongly associated with Criterion 8, which explains 82% of the variance. The problems are also similar to the those of the classification trees: The discriminants are strongly model dependent.

Finally, we have used logistic regression to distinguish between the true model and each of the misspecified versions. The identification of misspecification is largely successful, based on a subset of the model quality heuristics with some two-way and higher-order interaction effects. However, this method also suffers from the problem that results differ between models and also between different misspecification conditions. For example, the set of statistically significant criteria that distinguishes XL1 from the true model may not be the same as that which distinguishes XL2 from the true model. Further, even if the set significant predictors are identical, the parameters of the discriminant function will differ.

Discussion and Conclusion

We have examined the behavior of multiple model quality criteria for PLS under varying misspecification conditions using a simulation study. The criteria display a bewildering range of behavior, depending strongly, and in a complex way, on the model, type of misspecification, sample size, number of indicators, loadings, and strength of structural coefficients. We believe that many of the proposed evaluation and assessment criteria for PLS models are problematic for practical use because of their differential behavior under different conditions. For some criteria, there are widely varying values for the true models (e.g. for the GoF metrics, criterion 10) so that an absolute model assessment is problematic. Other criteria, such as the AVE relative to latent correlations (criterion 1), are problematic as they improve for the misspecified models. Still others, such as the item loadings relative to cross-loadings (criterion 8) are problematic because they are intended to identify measurement problems, yet are also or primarily sensitive to structural misspecifications. Finally, some criteria, such as the r^2 value (criterion 4) improve under some experimental conditions but deteriorate under others.

Second, we found many weak effects, i.e. comparatively small differences in means (less than 5% change) for different misspecification conditions. We believe this is problematic because, while the differences are statistically significant, they are sufficiently small that they might in practice be overlooked or dismissed by a researcher.For example, when the percentage of factor loadings >.7 is 100% in one model and 99% in another, it is unlikely that many researchers would consider such a difference to be indicative of model problems. In fact, the literature is replete with expressions like “all but two of our items satisfy condition X” or “construct validity *approached* condition X for most constructs”. Such approximations may in fact hide model misspecifications.

Given that PLS does not have an inherent ability to test models using a statistical test, but can only fit given models to data, we believe it is misleading to call PLS analyses exploratory. We acknowledge the term was coined to describe the fact that PLS estimates tend to overemphasize measurement loadings and under emphasize relationships between latent variables, thus emphasizing measurement in the presence of uncertain theory (Lohmöller, 1989). However, given that misspecified models are not likely to be rejected using PLS, at least not through the use of the proposed criteria, PLS appears to be a tool for fitting models of which researchers are very much certain about their structural relationships. This suggests that rather than exploring uncertain theory, researchers need a strong reason to believe in the basic model structure. This places an even greater onus on the researcher to ensure good theory development and internal validity, firmly basing theory on prior work or rigorous logic and ruling out alternative explanations. However, in many areas of IS research the relationships between constructs are ambiguous. For example, is the quality of a system judged to be high because it produces high-quality information, or is the information of high quality, because the system that produces it is a good system? Other areas of IS research are equally ambiguous. Ultimately however, suggesting that PLS should only be used when the theory is “solid” is of limited value, as it is precisely the theoretical “solidity” that is typically the subject of the research.

We also recommend that researchers use PLS for model comparisons. While PLS lacks the ability to reject wrong models, it is well suited to compare alternative models with respect to the various proposed evaluation or assessment criteria. For example, Chin (2010) and Goetz *et al.* (2010) suggest the use of an f^2 effect size and the Stone & Geisser test for comparing predictive ability of models with and without certain predictor variables of interest. When comparing models, the researcher needs to be explicit about conflicting aims. For example, one model may be better in terms of cross-loadings but worse in terms of goodness-of-fit or AVE. In such cases, the purpose and goal of the research can offer the context necessary to determine the “better” model. Such an argument should be explicit and

plausible. In this usage context of exploring alternative models, we agree that PLS can be called an exploratory technique, as it allows to explore different, plausible, theories.

In summary, we have shown that the model evaluation and assessment criteria proposed in the PLS literature may be suited for model comparisons, but we believe they are not suitable for identifying and rejecting misspecified models. Equally, our results indicate that the conformance of a PLS model to the proposed model quality criteria does not necessarily indicate a correctly specified model. This means that PLS models, even those which conform to widely used quality heuristics, are an uncertain foundation for theory building and testing.

However, as complex as the experimental design of this study is (1134 experimental conditions), there are issues that we did not examine. First, formative indicators are an area of concern that has seen much recent debate in the IS literature and related fields (Aguirre-Urreta and Marakas, 2008; Cenfetelli and Bassellier, 2009; Jarvis *et al*., 2003; Petter *et al*., 2007) and as such deserves more space than can be adequately given in this study. It is often argued that formative indicators are one of the strengths of PLS, hence the behavior of PLS estimates under various conditions of formative/reflective misspecification needs to be investigated and compared to, e.g. the findings of Jarvis *et al*. (2003) for covariance-based techniques. A second issue not examined here are higher-order factors in CFA models, which have recently gained some attention from PLS researchers (Wetzels *et al.*, 2009), although our Model 2 is a second-order factor model, although not of the form proposed by Wetzels *et al*. (2000). This omission is partially due to our desire to focus on models with endogenous latent variables as might be found in substantive, theoretical models. Hence, given the space limitations, higher-order factor models are left as a future extension of this research. Third, the inter-factor correlations could be examined as an experimental factor as they may have an impact on factor cross-loadings. In this study, the exogenous latent variables were defined as orthogonal while the endogenous latent correlations were the result of the respective regression coefficients, hence were indirectly changed with the changes in effect size *b*. For both, an explicit experimental factor could be introduced. Finally, as we indicated in our section on the study design, we did not the distribution of samples (their multivariate skewness and/or kurtosis), the response type (categorical/continuous), and the proportion of missing values. As we indicated in the section on study design, we believe that the chosen experimental factors are those typically used in simulation studies.

Returning to the notion of model fitting versus model testing, we suggested in the introduction that without indications of model correctness the parameter estimates (and their significance) become irrelevant as the model parameters do not correspond to any real or population parameters. In fact, while we have not reported on this in detail, we found that for all our experimental conditions all structural path estimates were statistically significant, despite the fact that the models were misspecified. This means that the PLS analysis indicates significant model parameters that have no equivalent real or population parameters. This makes the testing of causal relationships by means of PLS analysis problematic. The failure of the parameter significance tests to identify misspecifications is another reason, in addition to the failure of heuristics to identify misspecification, why we prefer to speak of model fitting instead of model testing, despite the fact that PLS does provide parameter tests. Moreover, this also shows that the set of parameter significance tests does not constitute and cannot substitute for an overall test of model correctness, such as is available in CB-SEM.

Finally, this paper has not addressed the question of whether model misspecification is a real problem. Given that PLS is unable to reliably identify misspecifications, it is perhaps better to speak of useful models, rather than true models. A PLS model is useful if it explains endogenous variances, as that is what PLS is designed to do. Predictive criteria, such as changes in Q^2 (q^2) or r^2 (f^2) proposed by (Chin, 2010) and Esposito-Vinzi et al. (2010) appear to be the most suitable indicators for a useful model. Whether or not a useful model is also the "true" model is irrelevant if we retreat to such a pragmatic philosophy, thus sidestepping the issue of misspecification altogether. However, we suspect that many IS researchers hold at least a somewhat realist position and would argue that science and research is a search for truth, and, in contrast to pragmatists, would make a distinction between true models and useful models. We believe two issues make such a pragmatist retreat problematic. First, the fact that CB-SEM is more reliable in identifying misspecifications suggests that a pragmatist retreat is premature and we need to continue developing further tests of model misspecification. Second, and more problematic, without understanding the true causal nature of a phenomenon, useful predictive tools such as PLS models can break down unexpectedly with possibly severe consequences.

To conclude, the usefulness of PLS lies in its ability to produce parameter estimates that maximize the predictive value of the fitted model. Thus, if the aim of the researcher is prediction, we recommend the use of PLS with appropriate predictive quality criteria. However, both researcher and readers must be aware that the models are merely fitted, but not tested.

References

Aguierre-Urreta, M.I. and Marakas, G.M. "The use of PLS when analyzing formative constructs: Theoretical analysis and results from simulations," *Proceedings of the 29th International Conference on Information Systems (ICIS)*, Paris, France, 2008

Areskoug, B. (1982). "The first canonical correlation: Theoretical PLS analysis and simulation experiments," in *Systems under Indirect Observation: Causality, Structure, Prediction*. K. Joreskog and H. Wold (eds.), Amsterdam, North Holland, 1982.

Cassel, C.M., Hackl, P. and Westlund, A.H. "Robustness of Partial Least Squares for estimating latent variable quality structures," *Journal of Applied Statistics*, (26**:**4), pp. 1999, 435-446.

Cenfetelli, R.T., and Bassellier, G. "Interpretation of formative measurement in information systems research," *MIS Quarterly*, (33:4) 2009, pp. 689-707.

Chin, W. W. "Issues and opinions on structural equation modeling," *MIS Quarterly*, (22:1), 1998.

Chin, W. W. (2010). "How to write up and report PLS analyses," in *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang (eds.), Springer Verlag, Heidelberg, Germany, 2010, pp. 655-690.

Chin, W. W. and Gopal, A. "Adoption intention in GSS: Relative importance of beliefs," *DATABASE*, (26:2), 1995, pp. 42–64.

Chin, W.W. and Newsted, P.R. "Structural equation modeling analysis with small samples using partial least squares," in: *Statistical Strategies for Small Sample Research,* Hoyle, R.A. (ed.), Thousand Oaks, CA: SAGE Publications, 1999.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., and Chen, F. "The noncentral Chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation," *Multivariate Behavioral Research*, (37:1), 2002, pp. 1–36.

Davey, A., Savla, J., and Luo, Z. "Issues in evaluating model fit with missing data," *Structural equation modeling*, (12:4), 2005, pp. 578–597.

Dolan, C., van der Sluis, S., and Grasman, R. "A note of normal theory power calculation in SEM with data missing completely at random," *Structural Equation Modeling*, (12:2), 2005, pp. 245–262.

Esposito Vinzi, V., Trinchera, L., and Amato, S. "PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement," In *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang, (eds.), Springer Verlag, Heidelberg, Germany, 2010, pp. 47–81.

Fan, X., Thompson, B., and Wang, L. "Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes," *Structural Equation Modeling*, (6:1), 1999, pp. 56–83.

Flora, D. B. and Curran, P. J. "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data," *Psychological Methods*, (9:4), 2004, pp. 466–491.

Fornell, C. and Larcker, D. F. "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, (18), 1981, pp. 39–50.

Gefen, D. and Straub, D. "A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example," *Communications fo the Association for Information Systems*, (16:5), 2005.

Gefen, D., Straub, D. W., and Boudreau, M.-C. "Structural equation modeling and regression: Guidelines for research practice," *Communications of the Association for Information Systems*, (4:7), 2000.

Goetz, O., Liehr-Gobbers, K., and Krafft, M. "Evaluation of structural equation moels using the partial least squares (PLS) approach," In *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang, (eds.), Springer Verlag, Heidelberg, Germany, 2010, pp. 661–711.

Gold, M. S., Bentler, P. M., and Kim, K. H. "A comparison of maximum-likelihood and asymptotically distribution-free methods of treating incomplete nonnormal data," *Structural Equation Modeling*, (10:1), 2003, pp. 47–79.

Goodhue, D., Lewis, W., and Thompson, R. "PLS, small sample size and statistical power in MIS research." In *Proceedings of the 39th Hawaii International Conference on Systems Science*, 2006, pp. 1–10.

Goodhue, D., Lewis, W., and Thompson, R. "Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators," *Information Systems Research*, (18:2), 2007, pp. 211–227.

Henseler, J., Ringle, C.M., and Sinkovics, R.R. "The use of partial least squares paths modeling in international marketing," *Advances in International Marketing*, (20), 2009, pp. 277-319,

Hulland, J. "Use of partial least squares (PLS) in strategic management research: A review of four recent studies," *Strategic Management Journal*, (20), 1999, 195–204.

Jarvis, D., MacKenzie, S., and Podsakoff, N. "A critical review of construct indicators and measurement model misspecification in marketing and consumer research," *Journal of Consumer Research,* (30:3) 2003, pp. 199-218.

Lei, M. and Lomax, R. G. "The effect of varying degrees of nonnormality in structural equation modeling," *Structural Equation Modeling*, (12:1), 2005, pp. 1–27.

Lohmöller, J.-B. *Latent Variable Path Modeling with Partial Least Squares*. Physica-Verlag, Heidelberg, Germany, 1989

Olsson, U. H., Foss, T., and Breivik, E. "Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central Chi-square distribution under model misspecification?" *Sociological Methods & Research*, (32:4), 2004, pp. 453–500.

Petter, S., Straub, D., and Rai, A. "Specifying formative constructs in information systems research" *MIS Quarterly*, 31:4), 2007, pp. 623-656.

Reinarts, W., Haenlein, M., and Henseler, J. "An empirical comparison of the efficacy of covariance-based and variance-based SEM," *International Journal of Research in Marketing*, (26), 2009, pp. 332–344.

Ripley, B.D. "*Pattern Recognition and Neural Networks",* Cambridge University Press, Cambridge, UK, 1996.

Rouse, A.C. and Corbitt, B. "There's SEM and "SEM": A Critique of the Use of PLS Regression in Information Systems Research", *Proceedings of the 19th Australasian Conference on Information Systems (ACIS), Christchurch, New Zealand,* 2008, pp. 845-855.

Savalei, V. and Bentler, P. "A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF," *Structural Equation Modeling*, (12:2), 2005, pp. 183–214.

Savelei, V. "Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data," *Structural Equation Modeling*, (15:1), 2008, pp. 1–22.

Straub, D., Boudreau, M.-C., and Gefen, D. "Validation guidelines for IS positivist research," *Communications of the Association for Information Systems*, (13:24), 2004.

Tenenhaus, M., Amato, S., and Esposito Vinzi, V. "A global goodness-of-fit index for PLS structural equation modelling." In: *Atta della XLII Riunione Scientifica, Bari, Italy, 9–11 June, 2004*.

Venables, W.N. and Ripley, B.D. *"Modern Applied Statistics with S*", 4th edition, Springer-Verlag, Berlin, 2002.

Wetzels, M., Odekerken-Schroder, G., and Oppen, C.v. "Using PLS path modeling for assessing hierarchical construct models: Guidelines and empirical illustration," *MIS Quarterly* (33:1) 2009, pp. 177-195.